Literature Review

1. “On improving wikipedia search using article quality” by Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw and Ba-Quy Vuong

Introduction: This section of the paper provides an overview of the search engine, including its indexing, ranking algorithms and its architecture. It also argued the limitations of these algorithms regarding the relevance and quality of search results.

Literature search: This section of the paper provides the various search engine ranking algorithms used in the literature including link-based algorithms, content-based algorithms, and keyword-based algorithms. It also discusses the challenges associated with the Wikipedia search engine, such as the large and dynamic nature of the content, the high variability in article quality, and the need for relevance and quality in search results.

Literature review: This section of the paper discusses the various standards of article quality used in the literature, including article length, number of citations, and number of revisions. The different embedding techniques used in the literature include word, sentence, and document embedding. It also discusses the limitations of these measures and the potential for incorporating other article quality standards, such as author reputation and accuracy of the content.

Synthesis and analysis: This section of the paper discusses the potential for incorporating user feedback into the search engine ranking algorithm, including user clicks, dwell time, and query reformulation. It also discusses the challenges associated with user feedback, such as user privacy concerns and potential bias. It also discusses the advantages and limitations of these techniques and the potential for using pre-trained embedding models such as BERT.

Conclusion: The paper concludes the various evaluation metrics used in the literature to evaluate search engine ranking algorithms, including precision, recall, and F1 score. It also discusses the limitations of these metrics and the potential for using user satisfaction and engagement as evaluation metrics.

2. “Leveraging BERT for extractive text summarization on lectures” by Derek Miller

Introduction: Text summarization plays an important role nowadays.Due to the excessive information available online, text summarization has become increasingly important.Extractive summarization methods, which select and combine important sentences from the original text, have gained popularity due to their simplicity and effectiveness. This paper discusses the use of BERT, a pre-trained transformer-based language model, for extractive text summarization in lectures.

Literature Review: The paper has discussed two studies on text summarization on lectures based on the BERT model.Gao et al(2020) proposed a BERT-based text summarization model for academic lectures.The model first encodes each sentence in the lecture using BERT model and search for the most informative text or sentences based on their contextual embeddings.Finally combined the most informative sentences to form the final summary.The author has evaluated model on the dataset of 30 lectures and compared it to several other models. Zhang et al(2020) proposed a BERT-based text summarization model for lectures.In this author has used the modified BERT model(LectureBERT) which was pre-trained on a large dataset of academic lectures.LectureBERT which had learned to capture the unique features of academic lectures and it is more effective the original BERT model.

Synthesis and Analysis: The two papers discussed the effectiveness of using BERT for text summarization on academic lectures.BERT model captures the informative sentences and key topics in lectures.BERT model produces the informative and concise summary.

Conclusion: The use of the BERT model for text summarization on lectures is the best approach for efficiently processing and retrieving information from academic lectures.The paper demonstrates the effectiveness of BERT models in capturing key topics in the lectures.

3. “A survey on word embedding techniques and semantic similarity for paraphrase identification” by Divesh R. Kubal and Anant V. Nimkar.

Introduction: Paraphrases mean the restatement of original sentences. This paper's purpose is that there is a problem to find the similarity between sentences or phrases that, if there is less lexical or syntactic overlap, then also point towards the same meaning.

Literature Review: Traditional techniques such as string-based similarity, corpus similarity, and knowledge-based similarity techniques are used. The traditional similarity measures can find similarities between pairs of sentences by considering word by word, not the whole sentence as one. So, the task of PI can be attempted by various techniques like SMT metrics, machine learning and deep learning techniques. A flow is maintained between the SMT, machine learning and deep learning techniques, which point towards deep learning algorithms to solve NLP problems.

Synthesis and Analysis: Our project uses the BERT model, a pre-trained language model that can be used to solve several NLP problems. There are some limitations in using N-gram, such as difficulty handling out-of-vocabulary words and limited context awareness, sparsity, and lack of semantic understanding; this is solved using the BERT model we will use in our project.

Conclusion: N-gram words in sentences should be considered to identify paraphrases in sentences. The traditional measures fail to consider n-grams. Supervised machine learning techniques solved this problem but to some extent only as they struggled to represent the data at multiple levels using deep learning based on representation learning. Using deep learning algorithms like

CNN and Recurrent neural networks outperform traditional methods' properties like sparse interactions, parameter sharing and equivariant representation. These three properties improved the accuracy of finding semantic similarity. The three types of granularity should be considered: word level, phrase level and sentence level.

4. “Sarcasm Detection in Tweets with BERT and GloVe Embeddings” by Akshay Khatri and Pranav P.

Introduction: Firstly the paper has defined what is the actual meaning of sarcasm i.e. basically defined as a sharp, bitter, cutting expression or remark and is sometimes ironic. Joshi et al., 2017 classified the sarcasm into three categories: rule based, deep learning based and statistical based. In this paper, the authors Akshay, Pravan and Anand had used the BERT and GloVe embeddings to find sarcasm in tweets.

Literature review: The paper has discussed several rule based approaches. Maynard and Green et al., 2014 used hashtag sentiment to identify sarcasm. Vaele and Hao et al., 2010 identify sarcasm in similes using Google searches to determine how likely a simile is. Riloff et al. (2013) look for a positive verb and a negative situation phrase in a sentence to detect sarcasm. In statistical sarcasm detection, most approaches use bags of words as features (Joshi et al., 2017). Some other features used in other papers include sarcastic patterns and punctuations (Tsur et al., 2010),

Synthesis and analysis: In this paper, authors have used the balanced sarcastic and non-sarcastic twitter dataset which consist of 2 fields: response and context. The words are of vector form having size 768 and generated for both the embedding techniques. When Bert fails, Glove comes into picture and uses unsupervised learning to generate a vector form of word. It is a non contextual embedding. GloVe embeddings were chosen to capture the overall meaning of a sentence in a smaller amount of memory. After embedding Classifiers like Linear Support Vector Classifier(LSVC), Logistic Regression(LR), Gaussian Naive Bayes and Random Forest were used. Scikit-learn is used to train data sets.

The analysis was done by using the metric F-measure. It is a measure of a test’s accuracy and is calculated as the weighted harmonic mean of the precision With Bert: A good result was seen from SVM and logistic regression results. Moreover, adding the context boosts the result. With GloVe: The results were better when compared to the BERT. Also, GloVe was much faster than BERT. The better results were given by logistic regression.

Conclusion: Word embeddings are an efficient tool for detecting sarcasm. They are incredibly helpful since they convey a word's meaning in a vector representation. Although GloVe provides the same vector for a word occurring numerous times it was better than Bert. Logistic regression consistently performed better than the other classifiers in this investigation.

5. “NLP based intelligent news search engine using information extraction from e-newspapers.” by Monisha Kanakaraj and S. Sowmya Kamath.

Introduction: This paper presents a customised news search engine that focuses on creating a database of news articles by utilizing effective text information extraction from a news website and from many e-news sources.

Literature search: This section provides methods for information gathering and extraction systems such as WordNet, a thesaurus of English language based on psycholinguist studies for matching the extracted content semantically to the title of the web page. TF-IDF is used for identifying the web page blocks carrying information relevant to the page's title. It also discusses the challenges associated with web page parsing as they are built using Content Management Systems. So, the web news page is generated dynamically and the web page is often not well formed due to this dynamically generated content.

Literature review: Different techniques were proposed for information gathering and extraction systems. Yang, Dingkui and Song, Jihua [1] proposed a technique to improve the extraction of the main content of the web page, ECON [2] method proposed by Yan Guo, Huifeng Tang, Linhai Song, Yu Wang. Oza, Alpa K and Mishra, Shailendra [3] proposed a data leaning technique.A Style Tree (ST) approach is proposed by M. Asfia, M. M. Pedram, and A. M. Rahmani [4].Zhou, Baoyao and Xiong, Yuhong and Liu, Wei [7] proposed an alternative approach to mine the information. Other approaches are class attribute based [12] methods.

Synthesis and Analysis: This section of paper discusses that considering the word count frequency alone cannot give much relevant results during news search as they lack semantics. Most of the current information extraction systems use word frequency and DOM tree approach to mine the information content of the web page. WordNet used for finding semantic relations among words has been studied .The accuracy has been increased when compared to the normal term frequency based method. The system’s functionalities can be broadly categorized into offline mode and online mode based on when various tasks are carried out. Web page crawling and storing of crawled data happens offline and user preference to query conversion, query retrieval and result presentation happens online.

Conclusion: Two separate experiments were conducted, firstly to evaluate the precision and secondly to measure recall. Based on the results, it can be observed that the semantic based proposed method outperforms the conventional frequency based method. As for future work we can improve the quality and relevance of the search results using ranking algorithms like Learning to Rank [11].